Federer

CUSTAT, THE CORNELL UNIVERSITY COMPUTER PROGRAMS IN STATISTICS

BU-206-M

S. R. Searle

Biometrics Unit and Cornell Computing Center

ABSTRACT

CUSTAT (Cornell University Statistics) is the acronym for the library of computer programs for statistical analyses that are available at the Cornell Computing Center. This paper provides short summaries of these programs, refers briefly to the procedures entailed in using them, discusses their establishment and future development and comments on some general problems associated with using such programs.

BU-206-M

CUSTAT, the CORNELL UNIVERSITY COMPUTER PROGRAMS IN STATISTICS

S. R. Searle

Biometrics Unit and Cornell Computing Center

University computing centers are frequently called upon to carry out the tedious arithmetic involved in the statistical analysis of data, particularly of large quantities of data. As a means of facilitating this, most computing centers have in recent years prepared libraries of ready-made or "canned" programs that can be used for a variety of problems. The programs of this nature available at the Cornell Computing Center are known by the acronym CUSTAT (Cornell University Statistics).

The Programs of CUSTAT

Thirteen programs for statistical procedures are complete and available. They are as follows.

MUREG	Multiple regression
FANOV	Factorial analyses of variance (factorial experiments)
CORMA	Correlation matrix (product moment correlations)
EQSOL	Equation solver (linear equations)
LATCE	Lattice designs (analyses of variance)
RANKO	Rank correlation analyses
XTABS	Cross-tabulations (2-way contingency tables)
KWAY	Cross-tabulations (up to 8 factors)
PROBT	Probit analysis
FACTAN	Factor analysis
REFAC	An amalgamation of MUREG, FACTAN and CORMA
FACTESSO	Factor analysis (adapted from a program written by the Esso Oil Corp.)
NORMA	Normal varimax analysis (a supplement to FACTAN)

Brief summaries of these programs are contained in the appendix.

General characteristics of each program are:

Ability to handle a wide variety of analyses within the framework of the program title.

Capability of handling both small and large amounts of data, on punched cards or magnetic tape.

Facility to analyze many variables simultaneously.

Opportunity for making transformations on data.

Availability of printed output of several kinds, including input and/or transformed data if required.

An example of these characteristics is that MUREG can handle up to 99,999 observations on each of 140 variables; from these, any number of linear regression analyses can be computed, by the user specifying for each, which of his variables is to be the dependent variable and which are to be the associated independent variables. Likewise FANOV can handle experiments of up to 8 factors and a total of nearly 20,000 levels, analyzing as many as 140 variables from the same experiment.

Using the Programs

The CUSTAT programs are available for anyone requiring them. Gaining access to them is a three-step procedure. First, the potential user must be aware of the general Computing Center procedures and charges, as detailed in "Cornell Computing Center, Facilities and Services", available from the Center. Second, the particular program that is to be used must be specified. The procedure for doing this is given in "CUSTAT, the Cornell Statistical System", also available at the Center. And thirdly, the particular details of the problem at hand must be specified in accordance with the documentation of the program being used. This too, can be had from the Center.

For each program in the CUSTAT collection, written material is available that describes in detail the operation of the program, the data it can handle, the calculations it can do and the output that can be obtained. This material is known as the program documentation. Operation of any program relies on preparing in addition to the data, a few punched cards that give to the program information about the data and about the calculations and output required. The documentation describes in detail just how these cards are to be prepared. Each documentation is self-contained, and although many of the programs have certain

operating characteristics in common with each other, each documentation is self-sufficient for the program it describes and no cross-referencing among documentations is needed.

The documentations are not statistics manuals. In some cases a few algebraic details are given, but in general the user is expected to be completely conversant with the statistical procedures involved and the assumptions underlying them. For two programs companion material is available: for PROBT there is "Notes on Probit Analysis", by S. R. Searle (BU-202-M in the Biometrics Unit Mimeograph Series) and for FACTAN there is "Introduction to Factor Analysis" by R. E. Granda. Both are available at the Computing Center.

The documentations do not discuss, in any way, the situations in which one statistical analysis or another should be used. Deciding which analysis, and hence which program, is appropriate to his data is entirely the user's problem. It is he who must decide which program is to be used; CUSTAT contains no decision-making process as to whether, for example, the appropriate analysis is that of regression or a split plot analysis. CUSTAT users needing occasional assistance in this regard can usually obtain it from one or other of the statistical groups to be found on campus in the School of Industrial and Labor Relations and in the Departments of Mathematics, Plant Breeding and Industrial Engineering. For example, the Biometrics Unit of the Plant Breeding Department offers daily consulting with qualified graduate students of statistics, available to anyone requiring assistance; as time allows and the need arises, faculty too are involved in this consulting service.

Although no guidance in matters statistical is given in the documentations of the CUSTAT programs they are not altogether brief, for care has been taken to describe accurately and clearly the various options available and the steps to be taken in preparing the necessary cards and data in order to get the required output. For each program the documentation concludes with data for a hypothetical example, description of the control cards needed to have them analyzed and a copy of the output as printed by the program.

Should any user of the CUSTAT programs have difficulty in understanding the

documentations, in preparing his work for the programs or in making them operate, he should refer to the CUSTAT consultant available daily at the Computing Center. This consultant gives assistance in the operation of the programs; he does not give advice on the appropriateness of one program over another.

Establishment of CUSTAT

The CUSTAT programming system was established in the spring of 1963 in an attempt to consolidate several earlier statistics programs developed from specific customer requests. Based on previous experience it was felt that statistics represented a sufficiently large portion of total computer usage as to warrant development of a unifying system. Such systems at other universities and research establishments had been claimed, in some instances, as highly successful, where in many cases programs are developed jointly by users and their computing center and modified to form part of a single system. Cornell therefore embarked on its own program. Doing so entailed one overriding question: for which statistical analyses were programs to be prepared and in what sequence? Analyses which past experience had shown to be in greatest demand, seemed to be as logical an answer as any, especially with limited programming resources available. On this basis CUSTAT was developed, with programs for multiple regression, factorial analyses of variance, and the product-moment correlation matrix taking first priority when the current computer, the Control Data Corporation 1604, was delivered in the fall of 1962.

Three years after initiation the CUSTAT system now contains thirteen programs, eleven of them of major importance. This number is considerably smaller than that in many comparable systems at other university computing centers. Undoubtedly this is due to the philosophy of preparing programs only according to demand. It is also due to the cost factor just alluded to, for the cost involved in terms of time, personnel and finance, in developing each of the programs must not be underrated. Each of them, with its many options for input, calculations and output, is no mean task to prepare. Recurrent testing during preparation requires numerous uses of the computer, after each of which further desk work is necessary for corrections and additional development. The Center has only limited personnel and computer time available each day for this work,

and with the large number of options in a program that should all be tested in various combinations with each other, the total elapsed time involved in preparing a program such as those in CUSTAT soon becomes appreciable. The complete development of eleven large programs in approximately three years is therefore very satisfactory, especially when for only one of those three years was there as much as one full-time programmer allocated to the task. At other times there was considerably less than one full-time programmer.

As already indicated, the programs in CUSTAT have taken some time to completely finalize. Since each program caters to numerous kinds of input data it is a practical impossibility to test them all before making practical use of the program. So, when a program is developed to the point of producing correct results in limited situations it becomes operational, and the ultimate testing then consists of using the program on varied forms of real data. Development of the program progresses during this time, including error detection and correction, and in due time it is declared complete. Even then it may not be entirely free of errors, for occasionally one will come to light when the program is used on data of some peculiar nature for which the program was certainly designed but never actually tested. Development is thus a dynamic process, with the perfect program being a limiting state. The occasional errors that arise very late in development are usually "obvious" in nature and easily spotted; errors that are not may, of course, remain unlocated, but it is doubtful if these are any more serious or more numerous than those that have been promulgated in the past from desk calculators.

Since developing a program for CUSTAT is a substantial task why, it might be asked, has use not been made of the statistics programs developed elsewhere. Part of the answer lies in the dynamic nature of the programming task, as just described. Programs written in other places are only useful when they have been fully documented. And frequently, by the time documentation would be available the program itself may have been improved and changed but with no corresponding changes made in the documentation. As a result, the documentation no longer describes the program it accompanies - and trouble in using the program can easily ensue. Furthermore, minor differences in computer specifications as between one installation and another sometimes give rise to difficulties in using

programs written at other places. Also, it sometimes happens that transcription errors arise in duplicating a program for transmittal. The program received is then in error, but the only indication of this may be that something goes wrong during operation. There will be nothing in the documentation as to what has gone wrong or why, and with no programmer available who knows the details of the program it will be a long, tedious and oft-time fruitless task to locate the error and correct it. A reason for not utilizing the extensive lists of programs that computer manufacturers often have available relates to uniformity. Such programs are usually written by a variety of people in different places and merely accumulated as a library by the machine manufacturer. As a result, operating procedures vary greatly from program to program. In contrast, the procedures for using the CUSTAT programs are all quite similar, a fact which facilitates their use and enhances their value. For all these reasons, CUSTAT has been developed. While many computing centers have had great success using programs written elsewhere, such as the BIMED series emanating from the University of California at Los Angeles, others have not had good fortune. Cornell therefore has its own system.

Customer usage of CUSTAT

From the fall of 1962 until the fall of 1964 preparation of CUSTAT programs was a part-time task for one programmer, much of whose energies had to be devoted to custom programming. By November 1964 three programs were completely finalized, with documentation available: MUREG (multiple regression), FANOV (factorial analyses) and CORMA (correlations). The use made of these programs by Computing Center customers from November 1964 through June 1965, the major part of an academic year, is shown in Tables 1 and 2. (Information on other programs as completed is also shown.) During eight months there were 1,444 jobs run on the three major programs, 60% of them on MUREG, 11% on FANOV and 20% on CORMA. Frequency of use, and the number of customers involved is shown for individual months in Table 1. The general increase in March is noticeable - corresponding, presumably, to completion of research projects for thesis work. The amount of time required for the jobs is summarized in Table 2. Only 46 hours were needed for 1,444 jobs of which by far the majority required less than

2 minutes. A further two hours were taken up by CUSTAT in the spring term, for the running of 212 jobs stemming from regular course work.

Assuming 21 working days per month the figures in Table 1 indicate approximately 6 jobs a day being run on CUSTAT each month, increasing to 12 in March and remaining at that level. This probably represents a worthwhile return in comparison to the total users' efforts on desk calculators had no such programs been available. At the current rate of charging for computer time, however, it does not represent a good return to the Computing Center for the overall cost of developing the programs.

Consequences of CUSTAT

The CUSTAT programs can be used by anyone who (with the necessary funds available) learns how to set up data for them. Unfortunately this is not always synonymous with having a sufficient understanding of the statistical methods involved to make appropriate use of them or even, in some cases, to plan appropriate analyses. The mere existence of the CUSTAT programs thus promotes the wrong use of statistics. Many instances could be cited: the customers who used analyses of variance on data more suited to regression; the client who made a regression analysis on data that included an arbitrary value for each of the many missing observations; the student who reproduced his data cards three times in order to have more data in a correlation analysis; and the student who analyzed just the means in a randomized block design, having thrown away the measurements on the individual plants, thirty of them in each mean. These are the serious misuses of statistics promoted by the existence of easy computing facilities. Less serious abuses also abound: the Ph.D. candidate, ten days before his thesis exam, who wanted the χ^2 value for each of 1,500 "contingency" tables, of order 2 x 2, the entries being means; the researcher who said "lets get everything" in using step-wise regression on fifteen variables; and the man who reduced all his variables except one to fractions of that one, "because everyone does it". Numerous examples of similar abuse could be quoted. These and many other erroneous uses of statistical analysis have certainly been perpetrated prior to the advent of high-speed computers, but many more are now likely to occur with the chore of tedious hours at a desk-calculator removed.

Indeed, the false uses may not only occur more often but may also be more serious, since they can now be made on very large amounts of data.

The obvious answer to the misuses of statistics that may arise from the existence of CUSTAT is to have appropriate consulting services associated with CUSTAT. As has been explained, these are available at both the student and faculty level. But customers cannot be compelled to use these services. Having, perhaps, used them for one analysis they proceed with others - and if the statistical methods involved are not appropriate for subsequent work they may never know it. Another situation can also arise. The customer coming to the Computing Center for consultation is seeking advice on how to get his calculations done. He is not looking for advice on what the calculations ought to be - and he sometimes resents being given it, and may well show his resentment. One cannot prevent him from using the computer to carry out his calculations, but tactful suggestion of alternative analyses is sometimes difficult.

These difficulties in the consulting task are really no different from those encountered in statistical consulting generally, but they are worth mentioning because they affect, in this instance, customers' attitudes to the Computing Center as a whole. For example, computing facilities were once blamed by a client for mistakes in his analyses because a line in an analysis of variance table was labeled "Error". "Look," he said, "the computer has made a mistake!"

The future of CUSTAT

Expansion of CUSTAT will depend very largely on resources available for programming staff. Within two or three years the present computer is likely to be supplanted by a new, larger machine, so in the interim the addition of new programs to CUSTAT is likely to be limited. Only three are envisaged; and even these may not be finalized:

POLYO	Polynomial regression
LEESQ	Least squares analysis of linear models
ANOVA	Analysis of variance for unbalanced data

Other programs may be requested from the Computing Center, but the chance of

their being written is slim unless very convincing evidence is produced of a widespread demand for them and unless also, support for the necessary work is forthcoming.

The largest programming task confronting CUSTAT is that of re-programming for a new computer. The cost factor involved is not slight; and if Tables 1 and 2 are any guide at all to the usage made of CUSTAT they give ground for debate as to the economies of CUSTAT and how they should be apportioned. For the total usage of 46 hours in 8 months barely substantiates, from the Computing Center's viewpoint, the justification made when initiating CUSTAT that statistics represents a substantial portion of total computer usage. To individual users of CUSTAT its existence is no doubt meritorious; but charging just one sector of the University, the Computing Center, with its development may well be questionable. Further information on the use made of individual programs is being accumulated, but there is little doubt that some are used considerably more than others. Programs used as much as MUREG will certainly be prepared again, but few of the others appear to have such wide appeal to the University as a whole. (In one case the Computing Center invited 25 people to a meeting to discuss development of a new program period, All were known, through prior activity, to be interested in the topic. Two of the 25 attended. The program is now part of CUSTAT, but it is difficult to believe there is widespread interest in it.)

Certain programs are of more interest to some people than to others. Thus PROBT (probit analysis) is of prime interest to entomologists; FANOV (factorial analyses) and LATCE (latice designs) are largely the concern of agriculturalists, biologists and home economists; FACTAN (factor analysis) and CORMA (correlations) are used mostly in the social sciences; and so on. Hence each program tends to be a more frequently used tool in some disciplines than in others. Resources for developing the tools might, therefore, be allocated accordingly. This could easily be achieved by departments, or groups of departments, having their own programming staff. Several sections of the University already operate in this manner, with great advantage to themselves. Not only do the large jobs get to the computer but the smaller ones do too, ones that may otherwise have remained undone. And because disciplines can hire programming personnel who

have some training and/or interest in appropriate subject-matter the communication problem between researcher and programmer is greatly reduced - a fact that facilitates computer usage enormously.

Such localization of programming leads to more satisfactory use of Computing Center facilities, for the Center is left freer to devote its resources to the task of providing top-rate machine, language, and systems service for all users - a task that is increasing in complexity as computers become larger, faster and more sophisticated. CUSTAT would be developed by co-operative effort between the Computing Center and its customers. While some programs were being prepared by the Computing Center others could be written in the departments interested in having them, with guidance and assistance from the Center to ensure successful incorporation into the CUSTAT system. At all times the Center would act as a clearing house and provide liason between users having common interests, to avoid duplication of effort. It would, as now, always provide assistance in matters concerned with feasability, computer technicalities and allied problems, including those relating to training of programmers and other personnel and to development of new techniques and new uses for the computer.

Sharing the programming task in this way would result in a sharing of the development charges. It would also ensure preparation of several programs simultaneously, in contrast to having them prepared more or less in sequence by the Computing Center. However, if this were no hardship to users of infrequently used programs (such as LATCE and PROBT for example), developmental costs could to some extent be apportioned to individual users by invoking a flat rate charge for each and every use of a CUSTAT program written by the Center. This would be additional to the computer charges based on time used. As of now, users of CUSTAT pay the same charges for computer time with a CUSTAT program, as they do with a program of their own. The additional charge for using a CUSTAT program could be in the nature of a programming fee, token indication of the time saved the user and in total, partial contribution to Computing Center resources for CUSTAT development. A fee in the order of $7.50 or $10 for each job run might be appropriate. One difficulty of such a scheme would be arranging it in a manner appropriate to necessary accounting procedures.

Customer usage of CUSTAT programs, November 1964 - June 1965

(Usage for teaching or Computing Center purposes is not included)

Table 1. Number of times used. (Number of people in parenthesis)

Program	Month (1964 - 1965) Nov.	Dec.	Jan.	Feb.	March	April	May	June	Total number of times used
MUREG	80 (24)	97 (29)	91 (31)	83 (22)	174 (47)	165 (38)	108 (38)	74 (26)	872
FANOV	27 (5)	23 (2)	22 (4)	23 (3)	15 (3)	12 (2)	29 (5)	9 (5)	160
CORMA	19 (6)	19 (4)	22 (7)	10 (7)	55 (10)	78 (11)	71 (15)	138 (8)	412
EQSOL	These programs were not on the CUSTAT system until April, 1965.					6 (6)	1 (1)		7
RANKO						2 (1)	6 (3)	4 (1)	12
Total times used	126	139	135	116	244	263	215	225	1463

Table 2. Summary of time used.

Program	Number of Times Used				Total time used (Hours)	Average time per usage
	Total	Percentage				
		Less than 2 mins.	2 mins. and less than 5	5 minutes and more		
MUREG	872	79%	13%	8%	23	1'35"
FANOV	160	60	20	20	$8\frac{3}{4}$	3'15"
CORMA	412	75	14	11	$14\frac{1}{2}$	2' 6"
				Total	$46\frac{1}{4}$	

APPENDIX: CUSTAT Program Summaries

MUREG

Multiple regression analysis.

This program carries out the standard calculations for multiple regression, for any number of observations on up to 140 variables. Input can be raw data or a correlation matrix. Output includes totals, means, corrected and uncorrected sums of squares and products, standard errors, the correlation matrix, standardized coefficients, regression coefficients and their standard errors, the multiple correlation coefficient, analysis of variance table for fitting the regression, the inverse of the matrix of corrected sums of squares and cross-products of the independent variables, and the residuals (differences between observed and predicted values of the dependent variable). Both direct and step-wise regression analyses are available.

FANOV

Factorial analyses of variance.

Standard analysis of variance procedures are carried out for data in any sort of factorial design (without missing data) of up to 8 factors, analyzing up to 140 variables simultaneously. Analysis for seven different designs are available, or the user can specify his own. Output consists of the input data, totals, and means, and the analysis of variance table showing sums of squares, mean squares, and F-values.

CORMA

Correlation matrix.

A correlation matrix can be computed for up to 85 variables, from an almost unlimited number of observations. The program is specifically designed to handle data having either several or many observations missing: correlations are computed from just the data available on each pair of variables. Output includes means, variances, covariance and standard errors of the two variables

in each correlation, and means, variances and standard errors of all observations on each variable. Corrected and uncorrected sums of squares and products are available, as well as the correlation matrix and a matrix of numbers of observations.

EQSOL

Equation solver.

This program inverts a series of matrices, of differing sizes (up to 100), and for each one solves up to 100 sets of equations. It is available as a self-contained program or as a FORTRAN subroutine.

LATCE

Lattice design analyses.

Standard analyses of variance are carried out for data obtained from experiments set up in any form of rectangular lattices, prime-powered resolvable one-restrictional lattices, or lattice squares. Output consists of tables of totals and means, analyses of variance showing sums of squares, mean squares and F-values, a list of adjustment factors and statistics derived from the analysis of variance and a table of observed and adjusted treatment means.

RANKO

Rank correlation analyses.

Data are ranked by this program, and rank correlation coefficients computed, for up to 140 variables. Output options are Kendall's tau with standard errors and confidence intervals, Spearman's rho, partial rank correlations and measures of concordance.

XTABS

Contingency tables.

This program produces row-by-column contingency tables. Up to 140 variables can be handled simultaneously, deriving as many tables based on any

two of them as the user requires. It allows for alphabetic recoding, grouping of variables and assigning of alphabetic names to variables. Output tables available are: of n's; of n's expressed as percentages of the grand total, of the row totals and of the column totals; of expected values of the n's; and of percentage contributions to chi-square. The latter is, of course, also calculated.

KWAY

Multi-classification cross tabulations.

This program carries out all the totalling necessary for the construction of multi-classification cross-tabulations tables, for 2 through 8 classifications per table. It can handle 140 classifications per observation, from which any number of tables may be obtained, each having 2, 3, 4, 5, 6, 7, or 8 classifications. Output is a series of tables containing every count and total required for construction of the desired table.

FACTAN

Factor analysis.

Standard factor analyses calculations are carried out. Input can be raw data or a correlation matrix. Output includes correlation matrix, covariance matrix, means and standard errors; the principal component solution and/or factor analysis solution, characteristic roots, characteristic column vectors, factor analysis communality estimates, factor matrix, residual matrices and factor scores.

REFAC

Regression, factor analysis and correlation.

This is a combination of programs MUREG, FACTAN and CORMA, suitable for data of limited quantity. It contains most output options of those programs, and it also includes the facility of calculating a correlation matrix from CORMA and using this directly as input to MUREG and/or FACTAN.

PROBT

Probit analysis.

Calculations are carried out for the probit method of analyzing quantal response data. Maximum likelihood estimates are obtained by an iterative procedure using weighed linear regression. Output after each iteration includes provisional and working probits: and after the last iteration summary statistics are given.

NORMA

Normal varimax analysis (a FORTRAN program).

This program can be used as a supplement to FACTAN. It carries out an orthogonal rotation of a matrix of factor loading (up to 85 variables and 20 factors) according to the Normal Varimax criterion. Output consists of factor loadings, components and variances, and communality estimates.

FACTESSO

Factor analysis, Esso, (a FORTRAN program).

This is an amended version of a factor analysis program written by the Esso Oil Company.